Uva Wellassa University

Regression and Design of an Experiments
Reflective Assignment

1. The researcher in the particular institute wants to identify the factors which affected the employee wages in that institute. He obtained the data from 30 employees in that institute and data includes the 7 variables (Promotions, Education, Foods, Job Security, Loyalty, Working Condition and Wages). Then he performed multiple regression analysis to find a suitable model for this relationship. The following partial Minitab output consist the results of that analysis.

Table 1: Regression Analysis: Wages versus Promotions, Food, ...

Predictor	Coef	SE Coef	T	P
Constant	-8.6400	9.429	-0.92	0.375
Promotions	0.4326	0.1861	2.32	0.036
Education	0.2013	0.1234	1.63	0.524
Food	-0.1113	0.2272	-0.49	0.632
JobSecurity	-3.3937	0.6938	-4.89	0.000
Loyalty	0.8005	0.4540	1.76	0.100
WorkingCondition	0.26878	0.07767	3.46	0.004

Table 2: Analysis of Variance(ANOVA)

Source	DF	SS	MS	F	P
Regression	.....	170522	.....	.....	0.000
Residual Error	.....	.....	.....		
Total	.....	174935			

a. What is the fitted regression model?

Multiple Regression Model

$Wages=B_0+B_1Promotions+B_2Education+B_3Food+B_4JobSecurity+B_5Loyalty+B_6Workingcondition$

$= B_0+ B_1X1+ B_2X2+ B_3X3+ B_4X4+B_5X5+ B_6X6$

$=-8.6400+0.4326X1+0.2013X2-0.1113X3-3.3937X4+0.8005X5+0.26878X$

b. Complete the ANOVA table in the above Minitab output (Table 2).

(DFR)=P-1=7-1=6

(DFE)=N-P=30-7=23

(DFT)=N-1=30-1=29

MSR=SSR/DFR
=170522/6

=28420.333

MSE=SSE/DFE

=4413/23

=191.8696

SSE=SSTotal-SSR
=174935
=4413

F=MSR/MSE
=28420.333/191.8696

=148.1232

c. Write down the relevant hypothesis to test the model significance and check the model significance using the corresponding p-value at 5% significance level and make the final decision.

H_0: $B_1= B_2= B_3=B_4=B_5=B_6=0$ Vs H_1=! 0 At least one B_i is not equal to 0

d. What are the significance estimated parameters in the model? Check them using the corresponding p-values at 95% significance level with suitable hypothesis and comment on your results (Table 1)

- **Promotions** => P=0.036

P<0.05

Therefore we reject H_0 and model is Significant .

- **Education** => P=0.524

P > 0.05

Therefore we do not reject H_0 and model is not Significant.

- **Food** => P=0.632

P > 0.05

Therefore we do not reject H_0 and model is not Significant.

- **JobSecurity** => P=0.000

P < 0.05

Therefore we reject H_0 and model is highly Significant.

- **Loyality** => P=0.100

P > 0.05

Therefore we do not reject H_0 and model is not Significant.

- **WorkingCondition**

=>P=0.004

P < 0.05

Therefore we reject H_0 and model is Significant

e. Comment the appropriateness of fitted model using Coefficient of determination value(R^2).

- R^2=SSR/SST

=170522/174935

=0.97477

This indicates that 97.5% of the variation in wages is explained by the independent variables included in model. This R^2 value is very close to 1 indicates points are lay close to a straight line

2. Four different coatings (A, B, C, & D) are being considered for corrosion protection of a certain type of metal pipe when it is buried in soil. To investigate whether the amount of corrosion depends on the coating, 15 pieces of pipe are available. Coatings enough for 5, 3, 4 and 3 pieces of pipe are available from type A, B, C, & D respectively. Each piece is to be coated with one of the four coating, and to be buried in soil for a fixed time, after which the amount of corrosion will be measured.

Suppose an engineer seeks your help to design an experiment and analyze data in order to determine whether all four coatings have the same capability to protect pipes from corrosion.

a. Explain clearly how an experiment should be carried out to test this corrosion protection of pipes

1. randomly assign each of the 15 pipes to one of four coatings (A, B, C, or D).

2. After coating, bury all pipes in the soil for the specified time. then measure the level of corrosion on each pipe.

b. What is the model that you will use?

one-way ANOVA model.

c. What is (are) the hypothesis (hypotheses) that you intend to test?

H_0: The mean amount of corrosion is the same for all four coatings (A, B, C, D).

H_1: At least one mean amount of corrosion differs from the others.

d. What is (are) the test statistic(s) that you will use to test above hypothesis (hypotheses)?

The F-statistic

Now suppose the engineer has carried out an experiment, and has recorded the data in a table as follows. The column C1, C2, C3 & C4 contain the amount of corrosion observed in pipes that were coated with coating A, B, C, & D respectively.

Amount of corrosion observed in pipes that were coated with coating A, B, C, & D respectively			
A	**B**	**C**	**D**
48	81	54	65
49	85	45	70
47	89	50	69
50		53	
53			

e. Suppose that in usual notation, SSE = 116.20, and SST= 3101.73. Construct the ANOVA table and interpret the results in ANOVA table with a suitable hypothesis

Source of Variation	SS	DF	MS	F
Regression	2985.53	3	995.177	94.205
Error	116.20	11	10.564	
Total	3101.73	14		

at significance level 0.05. $F_{0.05,3,11}=3.59$

$F>F_{0.05,3,11}$ (Mean that we can reject null Hypothesis)

f. Interpret the following MINITAB output in your own words by writing appropriate hypothesizes.

Table 04: All pairwise comparisons among levels of method

```
Grouping Information Using Tukey Method and 95.0% Confidence

Coating   N    Mean        Grouping
2         3   85.00         A
4         3   68.00           B
3         4   50.50               C
1         5   49.40               C

Means that do not share a letter are significantly different.
```

The MINITAB output provides information on pairwise comparisons among the levels of the coating method using the Tukey method and a 95.0% confidence level.

- Coating: Represents the different coatings being compared.
- N: Indicates the number of observations for each coating.
- Mean: Represents the mean amount of corrosion observed for each coating.
- Grouping: Assigns a letter (A, B, C, etc.) to each coating based on the results of thepairwise comparisons.

(H_0): The mean amount of corrosion is the same for all coatings.
(H_1): At least one pair of coatings has significantly different mean amounts of corrosion

the interpretation:

- - Coatings labeled with the same letter are not significantly different from each other in terms of their mean amount of corrosion.

- For example, coatings labeled with 'A' have a mean of 85.00 and are not significantly different from each other, coatings labeled with 'B' have a mean of 68.00 and are not significantly different from each other, coatings labeled with 'C' have means of 50.50 and 49.40 and are not significantly different from each other.

- However, the means with different letters are significantly different. For example, the mean for coating 'A' (85.00) is significantly different from the means for coatings 'B', 'C', and 'D'.

- Similarly, the mean for coating 'B' (68.00) is significantly different from the means for coatings 'A', 'C', and 'D'. The same logic applies to other pairwise comparisons.

- This output helps identify which coatings have significantly different corrosion protection capabilities.

3. An experiment to evaluate the effects of certain variables on soil erosion was performed on 20-foot-square plots of sloped tea land subjected to 2 inches of artificial rain applied over a 20-minute period. It has recorded the Soil Lost (in pounds/acre), Slop Gradient of the plot,

Length (in inches) of the largest opening of bare soil on any boundary and percentage of ground cover. The ANOVA table and the results of Regression analysis are as follows.

Table 01: Parameters Estimated

Predictor	Coefficient	SE coefficient	T	P
Constant	-1.88	18.13	-0.10	0.020
SG	77.33	44.51	1.74	0.006
LOBS	1.55	0.73	2.12	0.121
PGC	-23.90	13.43	-1.78	0.000

SG- Slop Gradient of the plot
LOBS- Length (in inches) of the largest opening of bare soil on any boundary
PGC- Percentage of ground cover

a. What is the fitted regression model?

Multiple regression Model

$$\text{Soil Lost}=B_0+B_1SG+B_2LOBS+B_3PGC$$

$$= B_0+ B_1X1+ B_2X2+ B_3X3$$

$$=-1.88+77.33X1+1.55X2-23.90X3$$

b. Complete the ANOVA Table given in below.

Table 02: Analysis of Variance (ANOVA)

Source	DF	Sum of Square Error	Mean Square Error	F value
Regression	3	659.79	219.93	95.6217
Residual	16	36.8	2.30	
Total	19	696.59		

c. Write down the relevant hypothesis to test the model significance.

H_0: $B_1= B_2= B_3=0$ Vs H_1=At least one B_i is not equal to 0

d. Check the model significance using the corresponding p-value at 95% significance level and make the final decision.

$F_{0.05,3,16}=3.24$

$F > F_{0.05,3,11}$

reject null hypothesis and model is significant.

e. What are the significance estimated parameters in the model? Check them using the corresponding p-values at 95% significance level with suitable hypothesis and comment on your results.

SG => P=0.006

$P < 0.05$ reject H_0 and model is Significant.

LOBS => P=0.121

$P > 0.05$ do not reject H_0 and model is not Significant.

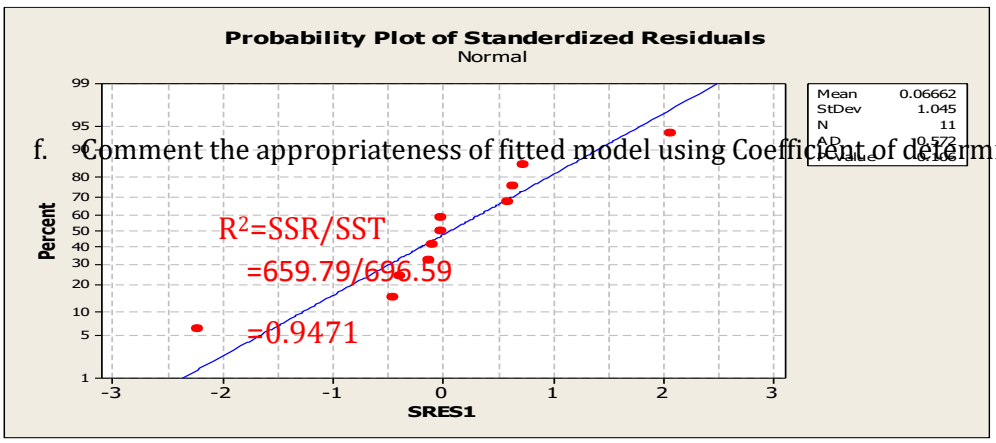
PGC => P=0.000

$P < 0.05$ reject H_0 and model is highly Significant.

e. The normal probability plot of standardized residuals of above analysis is given below. What comments can you make using Graph 01?

The points are closely near to the line,shows that the residuals are regularly distributed.

Graph 01: Probability Plot of Standardized Residuals



f. Comment the appropriateness of fitted model using Coefficient of determination value(R^2).

This indicates that 94.7% of the variation in soil lost is explained by the independent variables included in model. This R^2 value is very close to 1 indicates a better fit of model to the data and points lay close to the straight line.

4. A manufacturer of paper used for making grocery bags is interested in improving the tensile strength of the product. The Product Engineer thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentration s of practical interest is between 5% and 20%. A team of engineers responsible for the study decides to investigate 4 levels of hardwood concentration: 5%, 10%, 15% and 20%. They decide to makeup six test specimens at each concentration level, using a pilot plant. All 24 specimens are tested on a laboratory tensile tester, in random order. The data from this experiment are as follows.

Hardwood Concentration (%)	**Observation**					
	1	**2**	**3**	**4**	**5**	**6**
5	7	8	15	11	9	10
10	12	17	13	18	19	15
15	14	18	19	17	16	18
20	19	25	22	23	18	20

a. Suggest the suitable model to analyze this data.

Completely randomized design

b. Write down the appropriate hypothesis to test the significance difference in the hardwood concentration.

H_0: $B_1 = B_2 = B_3 = B_4$ Vs $B_i \neq B_j$ At least one pair is not equal

c. Obtain the degree of freedom for the Hardwood concentration, Error and Total and calculate the F-ratio value. (complete the below ANOVA table)

Source	**Degree of freedom**	**Sum of squared**	**Mean sum of squared**	**F-Ratio**
Hardwood concentration	3	382.79	127.60	19.6046
Error	20	130.17	6.5085	
Total	23	512.96		

d. Compare the calculated F-ratio value with F table value at 95% significance level and State whether each hardwood concentration is equally affected or not .

$F_{0.05,3,20} = 3.10$

$F > F_{0.05,3,11}$ reject null hypothesis

e. What are the model assumptions and mention appropriate tests to check the validity of assumptions?

- Normality: The residualsshould be normally distributed.
- Homogeneity of variances: The variances of the residuals should be approximately equal for all levels of the independent variable.
- Independence: Observations should be independent of each other.

Appropriate tests

- Normality: Shapiro-Wilk test, Kolmogorov-Smirnov test, or visual inspection of a Q-Qplot.
- Homogeneity of variances: Levene's test or Bartlett's test.
- Independence: This assumption is often assumed to be met unless there's a specificreason to believe otherwise.

5. A study was done to see which of four machines is fastest in performing a certain task. There are three operators; each performs the task twice on each machine. The MINITAB output follows.

Source	DF	SS	MS	F	P
Machine	(i)	257.678	(ii)	(iii)	0.021
Operator	(iv)	592.428	(v)	(vi)	0.000
Error	(Vii)	215.836	17.986		
Total	(Viii)	1096.646			

a. Select the appropriate design of the experiment for this study.

Randomized Complete Block Design

b. Write down the relevant linear model with the definitions of the term and statethe model assumptions.

$Y_{ij} = \mu+\tau_i+\beta_j+\epsilon_{ij}$

Y_{ij} represents the observation (performance time) for the machine and operator.
μ is the overall mean performance time.
τ_i is the effect of the i^{th} machine(i=1,2,3,4).
β_j is the effect of the j^{th} operator(j=1,2,3).
ϵ_{ij} is the random error term.

Assumptions,

- Normality: The residuals (differences between observed and predicted values) arenormally distributed.
- Independence: Observations are independent of each other.
- Homogeneity of variances: The variances of the residuals are approximately equalfor all combinations of machines and operators.

c. Fill the ANOVA table and compare the effectiveness of four machines.

Source	DF	SS	MS	F	P
Machine	3	257.678	85.892	2.388	0.021
Operator	4	592.428	148.107		0.000
Error	12	215.836	17.986		
Total	19	1096.646			

Assuming the significance level is 0.05.
PMachine=0.021 and POperator=0.00 -> Pm, Po <0.05
Reject Null Hypothesis

6. Suppose that a drug company wishes to test the effects of five new compounds on the growth rate of white rats. It is possible that rats within the same litter may have similar response. Twenty-five rats within each of 4 litters are chosen at random from a large group of rats, and 5 rats of each litter are placed in one pen to be given one of the five treatments. Body weight gains (g/day/animal) on the pen basis are given in the following table after 3 months of feeding the compound.

	Compound				
Litter	**1**	**2**	**3**	**4**	**5**
1	1.45	1.08	1.72	1.04	0.98
2	1.39	1.21	1.45	0.79	1.07
3	0.86	0.99	1.42	1.01	1.32
4	1.04	0.76	0.97	1.05	0.85

a. State the appropriate design of the experiment for this study.

Randomized Complete Block Design

b. Write down the relevant linear model with the definitions of the term.

$Y_{ij}= \mu+\tau_i+\beta_j+\epsilon_{ij}$

c. State the model assumptions.

- Normality: The residuals (differences between observed and predicted values) arenormally distributed.

- Independence: Observations are independent of each other.
- Homogeneity of variances: The variances of the residuals are approximately equal forall combinations of machines and operators.

d. Test the effects of five new compounds on the growth rate of white rats

Source of Variance	DF
Treatments	5-1=4
Block	4-1=3
Error	(5-1)*(4-1)=12
Total	20-1=19

7. A production manager has compared the scores of a test to check the fastness of employees in an assembly-line with their productivity. The data are shown in following table (Table 03).
Table 03: Data of Employees

Employee	**x = Score on test**	**y = Units produced in one hour**
A	12	55
B	14	63
C	17	67
D	16	70
E	11	51

a. To get a relationship between "score on test" and "units produced in one hour", what is the suitable analysis he has to perform?

Simple linear regression analysis.

b. Write down hypothesis for the particular analysis.

(H_0): There is no significant linear relationship between the score on the test and units produced in one hour

(H_1): There is a significant linear relationship between the score on the test and units produced in one hour .

c. Construct the complete ANOVA table for the test.

Source	DF	SS	MS	F
Regression	1	234	234	30.789
Error	3	22.8	7.6	
Total	4	256.8		

d. (Suppose that in usual notation, SSE = 22.8 , and SSTol = 256.8)

e. Check the model adequacy at 5% significance level (using F ratio).

$F_{0.05,1,3}$= 10.13

$F > F_{0.05,1,3}$

Reject Null hypothesis

f. Find the coefficient of determination (R^2) and interpret it.

R^2=SSR/SST

=234/256.8

=0.9112 (It is fitted linear Regression)

g. What are the model assumptions?

Linearity: There should be a linear relationship between the independent anddependent variables.

Independence: Observations should be independent of each other.

Homoscedasticity: The variance of the residuals should be constant across alllevels of the independent variable.

Normality: The residuals should be normally distributed with a mean of 0.

8. A multiple regression analysis is used to study the Sri Lankan tourist behaviour. A sample of 1000 respondents is considered. The objective of this study is to predict the number of miles travelled on vacation. The number of miles travelled on vacation, family size, family income and age are recorded. The description of variables is given in Table 01. Answer the following questions, in each case justifying your answer by appropriate analyses.

Table 01: List of variables

Variable	**Definition**
Y	number of miles travelled on vacation
X1	family size
X2	family income
X3	age

Using the computer software, the following incomplete Minitab output has generated for the best-filling multiple regression.

Table 02: Coefficients Table

Predictor	Coefficient	Standard Deviation	P
Constant	26.40	6.700	0.000
X1	-0.250	4.800	0.000
X2	5.02	4.690	0.123
X3	-0.269	0.587	0.000

Table 03: Analysis of Variance Table

Source	df	SS	MS	F
Regression		8925698		
Residual		132589		
Total				

a. What is/are the continuous variable/s?

Number of miles traveled on vacation
Family size
Family income
Age

b. What is the fitted regression model?

(Specify the null and alternative hypothesis, the test used in part c. and d. using a 5% level of significance)

Multiple linear regression model

Number of miles traveled on vacation= B_0+ B_1X1+ B_2X2+ B_3X3

=26.40-0.250X1+5.02X2-0.269X3

Hypothesis

H_0: B_1= B_2= B_3=0 Vs H_1=At least one B_i is not equal to 0

c. Test whether the multiple regression model is statistically significant, and explain what does it means.

F0.05,3,996 =2.61 F > F0.05,3,996 reject Null Hypothesis

It means at least one of the variables is a significant predictor of the number of miles traveled on vacation.

d. Which, (if any), of the independent variables are statistically significant? Explain your reasoning. What implications do these findings have?

X1=> P=0.000 P < 0.05 Reject Null hypothesis and model is highly significant.

X2=> P=0.123 P > 0.05 Do not reject null hypothesis and model is not significant.

X3=> P=0.000 P < 0.05 Reject null hypothesis and model is significant

e. How much proportion of variation can be explained by the fitted regression model?

R2=SSR/SST =8925698/8793109 =1.01

f. What are the model assumptions?

• Linearity: There is a linear relationship between the independent variables and the dependent variable.

• Independence: Observations should be independent of each other.

• Homoscedasticity: The variance of the residuals should be constant across all levels of the independent variables.

• Normality: The residuals should be normally distributed with a mean of 0.